Übungen zur Vorlesung Sequenzanalyse II

Universität Bielefeld, SoSe 2008
Prof. Dr. Jens Stoye · Dipl.-Inform. Peter Husemann

http://gi.cebitec.uni-bielefeld.de/teaching/2008summer/sequenzanalyse

Blatt 0 vom 11.04.2007
Abgabe am 18.04.2007 vor der Vorlesung

Aufgabe 1 Metriken:

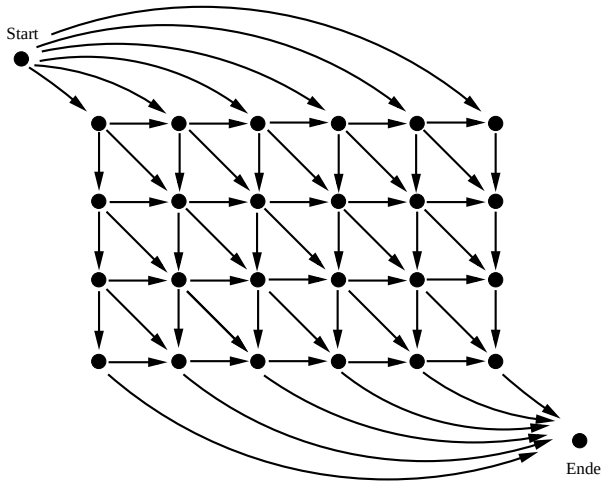
1. Gib die allgemeine Definition einer Metrik an.
2. Wie ist ein q-gram Profil definiert?
3. Wie ist die q-gram Distanz definiert?
4. Ist die q-gram Distanz $d_q : \Sigma^* \times \Sigma^* \to \mathbb{N}$ im allgemeinen eine Metrik auf Σ^*? Wenn ja, gib eine Begründung; wenn nein, zeige an einem Beispiel, welche Eigenschaft einer Metrik verletzt ist.

Aufgabe 2 Maximal-Matches-Distanz und Edit-Distanz:

1. Gegeben sei ein String $s \in \Sigma^n$ und $t := s\#s\#s$ mit $\# \notin \Sigma$. Berechne die Maximal-Matches-Distanzen $\delta(s\|t)$ und $\delta(t\|s)$.
2. Welche Schranke ergibt sich daraus für die Standard-Edit-Distanz $d(s,t)$?
3. Berechne $d(s,t)$ exakt.
4. Welchen (globalen) Edit-Score erhält man zwischen s und t, wenn man das Scoring-Schema $+1$ für einen match, 0 für einen mismatch und -0.5 für ein Indel verwendet? Wie ist der Zusammenhang zur Edit-Distanz?

Aufgabe 3 Alignment:

Hier ein Beispiel für einen schematischen Graphen für das glokale (andere Bezeichnungen: semi-global oder small-in-large) Alignment:



1. Zeichne den Graphen für das globale Alignment schematisch.
2. Zeichne den Graphen für das “free end gaps” Alignment schematisch.
3. Zeichne den Graphen für das lokale Alignment schematisch.

4. Wie sieht der Alignment-Graph $G = (V, E)$ für das globale Alignment aus? Gib genaue Definitionen der Knotenmenge V und der Kantenmenge E an.

5. Gib explizit die Rekursionsformeln (mit Initialisierung und Abschluss) im score-maximierenden Fall für das globale Alignment an. Jede verwendete Variable muss erklärt werden.

6. Wie werden in der Matrix für das lokale Alignment Start- und Endpunkt eines optimalen Alignments bestimmt?

Aufgabe 4 Suffixbaum und -array:

1. Zeichne den Suffixbaum von s = bcabcabc\$. Annotiere die Blätter mit den Suffix-Startpositionen und bilde so das Suffixarray `pos`. Beachte dabei die Ordnung \$ < a < b < c.

2. Bestimme das `lcp`-Array. Wie kann man es aus dem Baum ablesen?

3. Wie kann man in Linarzeit das "inverse" Suffixarray `rank` aus `pos` berechnen (Pseudocode)?

Aufgabe 5 Alignmentverfahren:

1. Wir wollen zwei bakterielle Genome der Größe jeweils 6 Megabasen mit Hilfe des Smith-Waterman-Algorithmus (lokales Alignment) vergleichen. Unser PC schafft es, eine Million Edit-Matrix-Einträge pro Sekunde zu berechnen. Pro Matrix-Eintrag speichern wir einen Score-Wert (`int`, 4 Bytes) und einen Backpointer (1 Byte). Wie viel Speicherplatz würden wir benötigen? Wie lange würde der Vergleich in Stunden dauern?

2. In einer der Übungsaufgaben zur Vorlesung Sequenzanalyse solltet ihr alle paarweisen FASTA-Scores $C(x, y)$ für fünf gegebene Proteine berechnen. Wie ist der FASTA-Score $C(x, y)$ für zwei Sequenzen x und y und einem q (Länge eines q-Grams) definiert? Falls ihr Probleme habt, die formale Definition aufzuschreiben, könnt ihr auch in Worten erklären wie der Score berechnet wird.

3. Wozu kann man den FASTA-Score benutzen? Welche Vor- und Nachteile bietet er?

Aufgabe 6 RNA-Faltung:

RNA-Sequenzen haben das Alphabet $\{A, C, G, U\}$ und die erlaubten Basenpaarungen A·U, U·A, C·G, G·C, G·U und U·G. Eine Struktur ist „erlaubt", wenn sie nur erlaubte Basenpaare hat.

1. Welche der vielen erlaubten Strukturen einer RNA-Sequenz x errechnet der Nussinov-Algorithmus? Ist das Ergebnis eindeutigt bestimmt?

2. Wieviele RNA-Sequenzen gibt es, die die folgenden Strukturen haben:
   a) `......`
   b) `((..))`
   c) `(.)(.)`

3. Ein Basenpaar in Position (i, j) heißt isoliert, wenn weder $(i - 1, j + 1)$ noch $(i + 1, j - 1)$ ein Basenpaar ist. Wieviele isolierte Basenpaare enthalten die Strukturen a), b) und c)?
   Gib eine kontext-freie Grammatik an, die alle RNA-Strukturen *ohne* isolierte Basenpaare erzeugt. (Hinweis: Die aus dem Skript bekannte Grammatik hat nur ein Nichtterminalsymbol – hier braucht man mehr als eines.)